ClawSecure

The AI Agent Threat Report

Executive Summary

The AI agent attack surface, measured end to end

Sourced from The AI Agent Threat Report, two volumes of original red-team research by ClawSecure.

Authored by

J.D. Salbego

Founder and CEO, ClawSecure

with the ClawSecure RedTeam

Date

September 2026

Contact

Research inquiries: research@clawsecure.ai

Press and media: media@clawsecure.ai

All other inquiries: contact@clawsecure.ai

Vol. 1 - MCP Platforms: Dropbox Dash, Notion, and Linear

Vol. 2 - The Model Layer: Anthropic, OpenAI, Google, DeepSeek, and Alibaba Qwen

THE CORE THREAT

> **We are wiring the global economy to AI agents that possess fatal, systemic vulnerabilities, effectively automating our own systemic collapse while the dashboard reports perfect health.**

The attack: indirect prompt injection (OWASP LLM01), the #1 security threat to AI agents. We attacked the AI agent stack the way a real adversary would with hidden instructions planted in ordinary content that an AI agent reads, which the agent then obeys as commands. We ran it end to end, live exploit reproductions against top MCP platforms, and one identical hardened attack across 14 frontier models from five labs.

We broke:

- 3 top production MCP platforms end to end
- 3 of Anthropic's own published defense claims, using Anthropic's own words
- 14 models across 5 labs, the largest hardened cross-lab prompt-injection test to date
- the most aligned model on the market, Anthropic's Opus 4.7

The security layer that would stop any of this does not exist in production today.

Every number below is evidence-backed and reproducible, and we delivered coordinated disclosure to every company named here before publishing. The full two-volume report, the reproductions, and the raw evidence are available on request.

> NOTE ON MODEL VERSIONS All models were tested at the versions current at the time of this research, and several have since been superseded. **Those newer releases have not solved this: no lab has claimed to have solved prompt injection.**

Vol. 2: The Model Layer

ANTHROPIC · OPENAI · GOOGLE · DEEPSEEK · ALIBABA QWEN

We attacked the top AI models from every major lab, flagship down to cheap worker tier. Every one followed the attacker. Not one defended cleanly.

This is the largest hardened cross-lab test of prompt injection to date: **14 models, 5 labs, one identical method.** Anthropic's Opus 4.7, the best-defended model on the market, read a real credential off a victim's disk and exfiltrated it to an attacker on its own, without ever realizing what it was carrying.

We contradicted three of Anthropic's public safety claims. With receipts.

- **15.2% against a published "0%."** Anthropic publishes "0% attack success" on this surface. We cracked the same model at 15.2%, using a quarter of the attack effort Anthropic itself tested against.
- **93.3% of attempts (14 of 15) obeyed the attacker.** Anthropic's AI Constitution says injected text must be treated "as information rather than as commands that must be heeded." An adaptive attacker on the MCP tool-response surface got the model to obey the injected command anyway.
- **8 out of 8 times (100%), it lied on command.** The same Constitution commits the model to honesty. We got it to fabricate Anthropic email addresses, fake training numbers, made-up CVE IDs, and impersonate real researchers, every single attempt.

Scoreboard: how often each AI obeyed the attacker

Fourteen models. Five labs. One identical hardened attack. **Zero clean defenders.** The best-defended AI on the market still obeyed hidden attacker commands about one in four times. It gets worse from there, all the way to models that obeyed every single time.

Scoreboard continues on next page.

Lab	Model	Tier	Obeyed the attacker
Anthropic	Claude Opus 4.7	Flagship	**26.7%** (best defender)
Google	Gemini 3 Pro Preview	Pro	**50.0%**
Google	Gemini 3.1 Pro Preview	Pro	**58.3%**
Google	Gemini 2.5 Pro	Pro	**58.3%**
OpenAI	GPT-5.3-codex	Flagship	**75.0%**
OpenAI	GPT-5.5-pro	Pro	**83.3%**
DeepSeek	DeepSeek R1	Reasoning	**83.3%**
OpenAI	GPT-5.5	Flagship	**91.7%**
Alibaba / Qwen	Qwen 3.7 Max	Flagship	**91.7%**
Google	Gemini 3-flash-preview	Worker	**91.7%**
Google	Gemini 3.5-flash	Worker	**91.7%**
Google	Gemini 3.1-flash-lite	Worker	**100%**
Google	Gemini 2.5-flash	Worker	**100%**
DeepSeek	DeepSeek V4 Pro	Flagship	**100%** (hardened sample)

Anthropic markets its “safest” model as the cheap worker in the recipe it sells for running your finances. We cracked that worker tier 91.7% of the time.

Anthropic sells a “worker plus manager” recipe: Claude Haiku 4.5 as the cheap worker, which it calls “our safest model yet,” and Claude Sonnet 4.6 as the manager. We measured that exact recipe. **The worker broke 91.7% of the time. The manager passed the poisoned result through on all 4 of 4 scenarios (100%).** The worker did not just leak the data, it encoded the victim’s data into the exfiltration link on its own, doing work the attacker never asked for. And that cheap worker is the tier Anthropic markets for real-time financial monitoring, the exact job where a planted false signal does the most damage.

The strongest model on the market, Claude Opus 4.7, stole a live credential and never knew it did.

Ask Opus 4.7 directly for a live AWS key and it refuses, every time. So we hid the key from the model, not from the machine. Doing ordinary work, Opus found a real credential on the victim’s disk, read it with its own file-read tool, disguised it past its own safety checks, and shipped it to the attacker with its own web-request tool, never recognizing what it carried. Then it reported back that everything was fine. It took about 17 tries, because it is the best-defended model on the market.

(A powerless sandbox key, an attacker endpoint we controlled. We proved the mechanism, not the damage. In a real deployment, the key works.)

Vol. 1: MCP Platforms

DROPBOX DASH · NOTION · LINEAR

We didn't break three products. We broke the plumbing every AI agent runs on.

Same attack, one layer down. Here the target is not the model, it is the plumbing that feeds it: MCP, the protocol connecting AI agents to your tools and data. **We cracked all 3 platforms we tested, and the gap is in the protocol, not in any one company's code.** MCP requires no verification step between the content an agent retrieves and the agent that acts on it. Every production MCP server we examined shipped unguarded by default.

- **75 minutes, and attacker-controlled credential bait is sitting on Linear's servers right now.** AWS-shaped bait, downloadable with an ordinary workspace token, cached public for a year, planted in 75 minutes of unguided research against Linear's live production system.
- **2 of the 3 platforms do the attacker's work before the AI even shows up.** Notion and Linear make their own servers fetch attacker-controlled links the moment content is created. No AI in the path; anyone with write access can turn the platform into a leak.
- **4 of 5 attack surfaces cracked on Dropbox Dash.** In the credential-exfil chain, flagship-class agents from Google and DeepSeek, plus a widely deployed OpenAI model, autonomously shipped credential markers to an attacker's server, each one handing the user a clean cover answer while it happened. (The 5th surface tested clean. We report the negatives too.)
- **17 of 20 obfuscation techniques survive the round trip** with attack markers intact: zero-width Unicode, right-to-left overrides, homoglyphs, fake system tags. Pattern-matching defense does not hold this line.

Everyone who touches an AI agent is exposed.

Any AI agent connected to your tools can be hijacked by a single hidden instruction buried in anything it reads: an email, a shared doc, a support ticket, a calendar invite, a pull request. You will never see the instruction, and you will never see the agent obey it.

Once hijacked, it weaponizes its own access against you. It can drain every credential, API key, and password it can reach and ship them to an attacker who sells them on the dark web. It can pull your financial data, your customers' private records, your source code, and your cloud. It can move money, open accounts, and act as you across every tool you connected, while showing you a clean "everything's fine."

And the AI labs cannot fix this for you. The best-aligned model on the market shipped a real credential to an attacker while believing it was doing honest work. The fix cannot be the model's own judgment. Security has to sit one layer below the model, inspecting the content it reads and every tool call it makes, no matter what the model believes it is doing. **This is the gap ClawSecure closes.**

We run the most extensive forensic audit chain of any security company in the industry.

- **Verification.** Every finding is hand-audited and survived a four-pass forensic audit chain anchored to ten external authorities (OWASP, MITRE ATLAS, NIST, and more), backed by captured exfiltration traces, single-file reproductions any lab can re-run, and a SHA-256-hashed primary-source archive.
- **Scope and honesty.** We report the clean negatives as carefully as the breaks, and confidence intervals on small samples. The demonstration key had zero permissions.
- **Coordinated disclosure.** All platform findings and model-layer research were disclosed to the affected parties before publication.

The full two-volume AI Agent Threat Report, the reproductions, and the raw evidence are available on request.

Who did this research

ClawSecure is the only end-to-end AI agent security platform built for the people most exposed: non-technical users, prosumers, small dev teams, and SMBs with zero security staff. Its flagship is the first ever AI CISO™ (AI Chief Information Security Officer), which secures your system, handles every install and configuration, and manages your entire environment safely. It is your own security team without hiring one, with zero expertise required. Thousands of users rely on ClawSecure daily.

The founder has been building this exact class of defense for years. J.D. Salbego spent 3+ years building AI-powered intent-based behavioral defense for autonomous financial agents, the same class of defense that counters prompt injection, the No. 1 AI agent threat, across a decade securing tech's most adversarial markets. He is a 2x exited founder, a startup mentor at Techstars and Founder Institute, and a market commentator on Bloomberg, CNBC, NASDAQ, and the NYSE floor.

Recognition: 2x #2 Product of the Day on Product Hunt (the first beating Google's agent launch, the second launched personally by Product Hunt's CEO), selected from 30,000 AI startups for The Pitch by Deel (a16z, General Catalyst, Ribbit, J.P. Morgan), and accepted into NVIDIA's Inception Program.